Dual Sparseness Constrained Nonnegative Matrix Factorization for Data Privacy and High Accuracy Utility

Bo PENG*, Xianfeng YANG

College of Computer Science, Southwest Petroleum University, Chengdu 610500, China

Abstract

In this paper, we propose a data distortion strategy based on dual sparseness constrained Nonnegative Matrix Factorization (NMF). The dual sparseness constrained nonnegative matrix factorization model incorporates attached term constrain and positive symmetric matrix into NMF, which is different from the previous approaches. The goal of our study is data perturbation and we study the distortion level of the algorithm with the standard NMF techniques and its sparse variants. K-means is used to evaluate the data utility of the proposed method. Experimental results indicate that, in comparison with previous published data distortion techniques, the proposed schemes are very promising solution for achieving both data privacy and data utility. At the same time, it provides a feasible method to protect sensitive information and promises higher accuracy for classification.

Keywords: Data Distortion; Data Mining; Privacy Preservation

1 Introduction

Privacy-preserving is a special multi-party collaborative computation and has many applications [1]. In recent years, many scholars have focused on privacy-preserving data mining (PPDM) problems. In collaborative data analysis fields of PPDM, a compromise between sensitive information for analysis and keeping each parties privacy and profit return motivated a great deal of research aimed to the increasing concern on PPDM. The general goal of PPDM is to hide sensitive individual data, and preserve the underlying data pattern simultaneously. According to this goal, our work is defined as to make sure that the distortion made to the data matrix has as less impact as possible to the utility of the data matrix. From a higher meaning, we wish the distorted data matrix is to provide better data analysis accuracy.

A variety of techniques have been studied by many researchers in this field. Most of these methods almost modify the values of the original datasets to protect confidential information. In
these methods, several randomization-based \cite{2,3} data distortion methods focus on perturbing the whole dataset using certain distribution of random noises like normal noise, uniform noise. Matrix decomposition and factorization, such as SVD, NMF, are put into use on original dataset for changing its value. Matrix decomposition has been traditionally used for the purpose of solving the linear system. In the PPDM field, the main goal of such methods has been in obtaining a simplified low-rank approximation version relative to original dataset. This kind of decomposition helps to study the structure of the data and it makes it easy to observe different relationship with the subjects and within attributes.

Distortion of dataset using SVD \cite{4} and NMF \cite{5} has been used for data privacy preservation and shown to be very promising in providing high level data privacy preservation and maintaining high accuracy data utilities. Recently, constrained nonnegative matrix factorization techniques \cite{6} have been also used to distort numerical valued datasets in the applications of PPDM.

In \cite{7}, Xu et al. proposed that NMF can be viewed as a kind of unsupervised learning in which the cluster label of each subject can be determined by the maximum value of each row of basis matrix. Inspired by this work, Wang et al. \cite{8} proposed a technique to simultaneously hide data values and confidential patterns without undesirable side effects on distorting non-confidential patterns. Our study is still based on the nature of basis matrix. From their work, we already know basis matrix may identify the cluster membership of each object in original dataset. In our experiments, we find that the accuracy of the clustering results by this rule is lower than that of the k-means clustering algorithm. However, this rule means that any modification on basis matrix may change the memberships of the corresponding objects. For this reason, on the contrary, if we try to keep the basis matrix and amplify the maximum value of each row of basis matrix, classification result will have a higher accuracy than the original dataset.

Basis matrix generated by the NMF represents additive combination for each subject which is the indicators for the cluster membership, while weight matrix represents coefficients for clusters \cite{8}. Based on this property of basis matrix, we intend to use sparseness constrained approach but this time for data distortion and not for explicit visual coding and representation.

In this paper, we investigate the standard nonnegative matrix factorization and its sparse variants. Based on previous works in \cite{4,5,6,8}, we propose a data distortion strategy based on dual sparseness constrained nonnegative matrix factorization (dsNMF) for PPDM. The basic idea underlying the proposed strategies is to force ultimate sparseness constraints on weight matrix and maximize basis matrix. Maximized basis matrix implies an underlying fact that reconstruction matrix may provide a higher classification accuracy.

This paper is organized as follows. In Section 2, we review the traditional standard nonnegative matrix factorization techniques and its partial variants. In Section 3, we present our dual sparseness constrained nonnegative matrix factorization for PPDM. Section 4 offers a brief description of data distortion and utility measures. Experiments and testing results of our method from different dataset are shown in Section 5. Conclusions are given in Section 6.

## 2 Background and Related Work

In this section, we will briefly describe PPDM based on NMF method and some of the extending works that, to the best of our knowledge, are closely related to the work you can find in \cite{6,8,9}. In the addition, two NMF variants are induced here.
2.1 Nonnegative matrix factorization (NMF) for PPDM

Nonnegative matrix factorization (NMF) [10] has been introduced as a matrix factorization technique that produces a useful decomposition in the analysis of data. NMF decomposes the data as a product of two matrices that are constrained by having nonnegative elements. The two matrices called basis and weight matrix. This method results in an approximate representation of the original data that can be seen as a data distortion technique in PPDM.

Given a nonnegative matrix $A \in \mathbb{R}^{n \times m}$ with $A(i, j) \geq 0$ and a pre-specified positive integer $k < \min(n, m)$, NMF finds basis matrix $W \in \mathbb{R}^{n \times k}$ with $W(i, j) \geq 0$ and weight matrix $H \in \mathbb{R}^{k \times m}$ with $H(i, j) \geq 0$ so that $A = WH$ that minimizes the objective function. The usual way to find $W$ and $H$ is by minimizing the Euclidean distance $\|A - WH\|$. The update rules based on the Euclidean distance:

$$H_{ab} \leftarrow H_{ab} \frac{(W^T V)_{ab}}{(WTH)_{ab}}$$  \hspace{1cm} (1)

$$W_{ia} \leftarrow W_{ia} \frac{(VHT)_{ia}}{(WHH^T)_{ia}}$$  \hspace{1cm} (2)

Another useful measure is Kullback-Leibler divergence [11]. The objective function based on the Poisson likelihood, is:

$$D(A, WH) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left( A_{ij} \ln \frac{A_{ij}}{(WH)_{ij}} - A_{ij} + (WH)_{ij} \right)$$  \hspace{1cm} (3)

The update rules based on the Euclidean distance:

$$H_{ab} \leftarrow H_{ab} \frac{\sum_i W_{ia} V_{ip} / (WH)_{ip}}{\sum_k W_{ka}}$$  \hspace{1cm} (4)

$$W_{ia} \leftarrow W_{ia} \frac{\sum_p H_{ap} V_{ip} / (WH)_{ip}}{\sum_H H_{av}}$$  \hspace{1cm} (5)

2.2 Sparse nonnegative matrix factorization

Liu et al. [11] modified the method described previously [12]. They used a divergence term instead of using an Euclidean distance in the original NMF. Thus, the sparse NMF functional is:

$$D(A, WH) = \sum_{i=1}^{p} \sum_{j=1}^{n} \left( A_{ij} \ln \frac{A_{ij}}{(WH)_{ij}} - A_{ij} + (WH)_{ij} \right) + \alpha \sum_{i,j} H_{ij},$$  \hspace{1cm} (6)

for $\alpha \geq 0$, This method forces sparseness via minimizing the sum of all $H_{ij}$. The update rule for matrix $H$ is:

$$H_{ab} \leftarrow H_{ab} \frac{\sum_{i=1}^{p} (W_{ia} A_{ib}) / \sum_{k=1}^{q} W_{ik} H_{kb}}{1 + \alpha}$$  \hspace{1cm} (7)
2.3 Nonsmooth nonnegative matrix factorization (nsNMF)

Inspired by the various sparseness NMF methods, Pascual-Montano et al. [13] present a novel Nonnegative Matrix Factorization algorithm intended for controlling sparseness of both basis and encoding parts using a smoothing matrix. The new model proposed in this study, denoted as "NonSmooth Nonnegative Matrix Factorization" (nsNMF), is defined as:

\[ A = WSH, \]  

(8)

where \( A \), \( W \) and \( S \) are the same as in the original NMF model. The positive symmetric matrix \( S \in R^{k \times k} \) is a "smoothing" matrix defined as:

\[ S = (1 - \beta)I + \frac{\beta}{k}11^T, \]  

(9)

where \( I \) is the identity matrix, \( 1 \) is a vector of ones, and the parameter \( \beta \) satisfies \( 0 \leq \beta \leq 1 \). Note that the parameter \( \beta \) controls the extent of smoothness of the matrix operator \( S \). This method forces sparseness constrained via a positive symmetric matrix. Due precisely the simultaneity of both conditions, sparseness will be enforced on both basis and weight matrix. However, the level of sparseness of basis and weight matrix is different. The update rule for matrix is a simple modification of the original, in the update equations for \( H \) (4) and for \( W \) (5), substitute \( W \) with \((WS)\) and \( H \) with \((HS)\) respectively.

In comparison with the direct modification of the model as the means to achieve sparseness, the critical aspect of nsNMF method is to achieve sparseness by using a positive symmetric matrix.

3 Dual Sparseness Constrained Nonnegative Matrix Factorization

Because of the multiplicative nature of update rule of NMF, i.e., “basis” multiplied by “weight,” sparseness in one of the factors will almost certainly force “nonsparseness” or smoothness in the other, in order to compensate for the final product to reproduce the data as best as possible. According to such multiplicative nature, if we force \( W \) (basis matrix) sparse, then \( H \) (weight matrix) will inevitably lead to non-sparse, that means smooth effect. In the aforementioned description, the maximum element of each row of \( W \) may denote each object cluster. If a constrained process will smooth \( W \) matrix, resulting from forcing \( H \) sparseness, therefore, reconstruction matrix not only improves the accuracy of classification, but also reaches the purpose of protecting sensitive information.

The proposed techniques use dual constrained techniques. First constrained method forces sparseness via minimizing the sum of all \( H \) (weight matrix), which derived from Sparse Nonnegative Matrix Factorization (SNMF). Second method applies a positive symmetric matrix to control the sparseness of \( W \) and \( H \). This idea have motivated the modification of original NMF and append positive symmetric matrix as the means to achieve the certain level sparseness about \( H \), which is different from previously published methods. If we adjust \( S \) matrix and the attached constrained term, we are able to have the result we want to get.
Thus, the sparse NMF functional is:

\[ D(A, WSH) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left( A_{ij} \ln \frac{A_{ij}}{(WSH)_{ij}} - A_{ij} + (WSH)_{ij} \right) + \alpha \sum_{ij} H_{ij}, \]

where \( \alpha \geq 0 \). The update rule is similar to nsNMF, we only need substitute \( W \) with \((WS)\) and \( H \) with \((HS)\) in the corresponding equations of SNMF. However, during the reconstruction step, we omit positive symmetric matrix and reconstruct the distorted matrix using \( W \) and \( H \) for better classification accuracy and data preserving.

### 4 Data Distortion Measures

We choose the three data distortion privacy measure metrics, VD, RP, and RK, first defined in [14], to evaluate the proposed data distortion methods. These measure metrics are to indicate how closely the original data values can be estimated from the counterpart distorted data and how much they are distorted. In brief, VD of the datasets is represented by the relative value difference in the Frobenius norm. RP denotes the average change of rank for all the attributes. RK represents the percentage of elements that keep their ranks of magnitude in each column after the distortion. According to their definitions, we know that a larger VD and RP value and a smaller RK value refer to a better privacy-preserving level.

### 5 Data Utility Measure

Data utility measures indicate the accuracy of data mining algorithms on distorted data after the conduction of certain perturbation. In this paper, K-means classification is chosen as the data utility measure. K-means is one of the most popular clustering algorithms among many. During these experiments, we treat original clusters of these dataset as ground truth.

### 6 Experiments and Results

We conduct some experiments to verify the performance of the data distortion methods: ND, SVD, NMF, SNMF, nsNMF and the proposed method. Experiments were performed with IRIS and YEAST datasets [15], both of which are fairly known dataset.

#### 6.1 Experiments on IRIS dataset

IRIS is a simple data set with 150 instances in a 4-dimensional attribute space. The four attributes are sepal length, sepal width, petal length and width. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant: Iris Setosa, Iris Versicolour, and Iris Virginica. Table 1 shows comparison of the distortion measures and the utility measure of applying these data distortion methods. The rank k in SVD is 3. NMF’s algorithm is normalized Kullback-Leibler divergence. The dual sparseness NMF described in the previous section are applied in the experiment, represent by dsNMF in Table 1. It is found that, compared with
the other data distortion methods, the proposed method work very well and provide comparable levels of data privacy and data utilities.

<table>
<thead>
<tr>
<th>Level of Distortion</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VD</td>
<td>89.3333</td>
</tr>
<tr>
<td>RP</td>
<td>74.6667</td>
</tr>
<tr>
<td>RM</td>
<td>61.4583</td>
</tr>
</tbody>
</table>

6.2 Experiments on YEAST dataset

YEAST is a real-valued data set having 1484 instances and 8 attributes. It is used to predict the localization site of protein, which has 9 classes. We performed some modifications on the original YEAST to make it suitable for our tests, which include:

1. Choose three classes of data from YEAST dataset, ME3, ME2, ME1 respectively.
2. Substitute zero value with a minimum value in the truncated dataset.

Similar to the experiments on IRIS dataset, a comparison is conducted on the YEAST dataset. The results of performance evaluation on the seven methods are provided in Table 2.

<table>
<thead>
<tr>
<th>Level of Distortion</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VD</td>
<td>58.6806</td>
</tr>
<tr>
<td>RP</td>
<td>30.2083</td>
</tr>
<tr>
<td>RM</td>
<td>61.4583</td>
</tr>
</tbody>
</table>

For the truncated YEAST dataset, the rank k in SVD is 3. NMF algorithm is also normalized Kullback-Leibler divergence. From Table 2, we can see that, dsNMF provide a level of data distortion and classification accuracy comparable with the other data distortion methods.
6.3 Experiments on parameter choosen

This experiment is to study the relation between the value of alpha and beta in S matrix. We assume that the sum of alpha and beta equals 1. Then, we performed two experiments both with the IRIS and the YEAST data for checking the accuracy underlying different the ratio of alpha and beta. For each dataset, experiment was repeated with the initial value of alpha = 0.1 and beta = 0.9 and increase the value of alpha by 0.1 and decrease beta by 0.1. In order to get the best accuracy, we set the max iteration number as 3000 and tolerance=10E-4. It can be seen from the two figures that better clustering accuracy can be achieved while the ratio of alpha and beta is close to 1.
7 Conclusion

In this paper, we proposed a novel privacy preserving data distortion methods based on sparseness constrained strategies of NMF for collaborative data analysis. Through experiments, we demonstrate that the proposed approach can perform much better than other methods in terms of clustering accuracy and preserving data privacy in collaborative data analysis. In the following work, we will combine other constrained conditions with the proposed method for better result.

References